VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MODELLING AND DATA ANALYSIS MASTER'S STUDY PROGRAMME

Master's thesis

Outlier Detection in Contingency Tables

Išskirčių aptikimas dažnių lentelėse

Anastasija Jeriomova

Supervisor: assoc. prof. Viktor Skorniakov

Vilnius, 2021

Outlier Detection in Contingency Tables

Abstract

In this work a new method for outlier detection in contingency tables is being examined, which is proposed in one of the most recent works in this field. The advantage of this method is that it is suitable not only for two-dimensional, but also for high-dimensional contingency tables. To our best knowledge, this type of method has not been proposed so far.

The main goal of the thesis - by making use of Monte-Carlo modeling to investigate operational characteristics of the suggested method and to evaluate how these characteristics vary depending on table dimensions, the magnitude of deviation of outliers, the structural model used.

The work provides with a brief overview of the method, a description of the modeling plan and the results obtained.

Key words : outliers, contingency tables, graphical models, sensitivity and specificity, Monte-Carlo modeling

Išskirčių aptikimas dažnių lentelėse

Santrauka

Darbe nagrinėjamas naujas dažnių lentelėms skirtas išskirčių aptikimo metodas, pasiūlytas viename naujausių šios srities darbų. Šio metodo privalumas yra tas, kad jis tinka ne tik dvimatėms, bet ir daugiamatėms didelių matmenų dažnių lentelėms. Tokio tipo metodų mūsų žiniomis iki šiol pasiūlyta nebuvo.

Pagrindinis darbo tikslas – pasitelkiant Monte-Karlo modeliavimą ištirti pasiūlyto metodo operacines charakteristikas ir įvertinti kaip jos kinta priklausomai nuo lentelių matmenų, naudojamo struktūrinio modelio, išskirčių nuokrypio didumo.

Darbe pateikiama trumpa metodo apžvalga, modeliavimo plano ir gautų rezultatų aprašymas.

Raktiniai žodžiai : išskirtys, dažnių lentelės, grafiniai modeliai, jautrumas ir specifiškumas, Monte-Karlo modeliavimas

Contents

1 Introduction

Outlyingness plays an important role in statistical inference since outliers may affect quantitative analysis quite significantly, changing this way overall qualitative conclusions. Though intuitively clear, the concept of outlying observation does not have general and broad enough quantitative operational definition and, as a rule, is tied not only to the nature of the observed data (discrete, continuous, univariate, multivariate, etc.) but also to the assumed data generating model. This applies to the tables of frequencies as well.

There are plenty of works and articles devoted to the analysis of outliers' detection in contingency tables. Several articles, including the paper we mainly analyse in our thesis, were used to get familiarized with the subject: Lindskou [1], Kuhnt [2], Rapallo [3], Yick [4]. Consequently, it is not surprising that in the literature devoted to categorical data analysis one can find several definitions of contingency tables' outliers and that none of these can handle all situations encountered in practice.

Although the literature on outlier detection is vast, the problem of detecting outliers in contingency tables has mainly been focused on two-way tables. No example was given for tables with dimensions larger than three. In [1], authors suggest a novel outlier detection method suitable for high-dimensional contingency tables. Since nowadays the amount of accumulated data increases very quickly, such methods play an important role and deserve attention.

In our work we mainly concentrate on this new method suggested by the authors of [1], as the goodness of its performance is curious to us. The novelty of the suggested method is that it must be suitable not only for two-way contingency tables, but also for high-dimensional tables as was indicated above. There are lots of methods for anomaly detection in low-dimensional contingency tables, but the method we are going to investigate in this thesis is the only method for now appropriate to detect outliers in high-dimensional contingency tables.

The main aim of the Thesis is to investigate the robustness of the newly suggested method of [1] article. To this end, we seek to give an answer to the question "how dimensionality of the table affects the performance of the suggested method assuming that data generating mechanism and outlyingness definition fully complies with those presumed in the paper [1]?". In order to answer this question, we conduct a simulational study where, by making use of Monte-Carlo method and for varying table dimensions and structure, we estimate whether dimensionality of the table affects such operational characteristics of the method as sensitivity and specificity. To be more precise, does the method identify outliers equally well (or bad) for tables having different dimensions or there are performance differences between low dimensional and high dimensional tables? We are also investigating the relationship between operational characteristics and table sparsity assuming, again, that all assumptions fully comply with those stated in the paper.

Another question is related to method's performance when some of the model assumptions under which it was developed are violated. Does the method perform well when not all assumptions of the model are met?

The structure of the thesis is as follows:

1. Literature review, description and explanation of the model and terms
2. Description of the simulation plan
3. Results
4. Discussion and main conclusions

2 Literature review and description of the model

In Kuhnt work "Outlier detection in contingency tables based on minimal patterns" [2] a new algorithm was developed to identify outliers in contingency tables. This algorithm is based on definition of minimal patterns, which are subsets of cell counts. The proposed algorithm, however, is suitable only for two-way contingency tables.

Rapallo [3] in his article defines the notions of outliers and its patterns by making use of log-linear models and goodness-of-fit tests. For making a definition more clear, author invokes techniques from algebraic statistics.

Yick and other authors of [4] propose an iterative testing procedure coupled with perturbation diagnostics for confirming multiple outliers in two-way contingency tables.

In [1], authors present a novel outlier detection method for high-dimensional contingency tables. They use the class of decomposable graphical models to model the relationship among the variables of interest, which can be depicted by an undirected graph called the interaction graph. Having an interaction graph, authors derive a closed-form expression of the likelihood ratio test statistic and an exact distribution for efficient simulation of the test statistic.

Authors of [1] paper focus mainly on high-dimensional tables and the application of the method in forensic genetics. They demonstrate the use of the LRT outlier detection framework on genetic data modeled by Chow–Liu trees. However, the method described in the article is general and applies to any outlier detection problem in contingency tables including sparse tables as well.

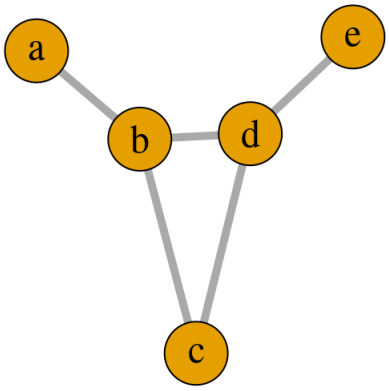
Since we mainly focus on the paper [1], we provide a detailed description of the model presented in the article below.

2.1 Preliminaries on graphs

A graphical model is a statistical model for which an undirected graph represents the interaction between the vertices in the model. The class of decomposable graphical models is used in [1] to model the relationship among the variables of interest, which can be depicted by an undirected graph called the *interaction graph*.

An undirected graph is a pair $G = (\Delta, E)$ where Δ is a set of vertices and E is a set of edges connecting elements in Δ. An edge connecting two vertices indicates that these two are dependend on each other and this is called a two-way interaction. A three-way interaction occurs when three vertices are all mutually connected.

Figure 1: G: An undirected decomposable graph



An undirected graph is *decomposable* if there are no cycles of length greater than four without a chord, which is an edge between two non-adjacent vertices in the cycle.

The subgraph $G_A = (A, E_A)$ consists of vertices $A \subseteq \Delta$ from G and the corresponding edges E_A between these vertices.

A graph is *complete* if there is an edge between all pairs of vertices.

A complete subgraph is called a *clique* if it is not contained in any other complete subgraph. A subset of vertices is complete if it induces a complete subgraph.

Two sets $A, B \subseteq \Delta$ are separated by a third set $C \subseteq \Delta$ if all paths between vertices in A and B go through C. If C is the smallest set such that A and B are separated, we say that C is a separator for A and B.

Consider the undirected graph G in the Figure 1. The set of vertices is $\Delta = \{a, b, c, d, e\}$ and the set of edges is $E = \{ab, bc, bd, cd, de\}$. The cliques are $C_1 = \{a, b\}$, $C_2 = \{b, c, d\}$ and $C_3 = \{d, e\}$. The minimal separators are $S_2 = \{b\}$ and $S_3 = \{d\}$ where S_2 separates C_1 and C_2 and S_3 separates C_2 and C_3. Since G has no cycles of length greater than three, the graph is decomposable. A *decomposable graph* is one that can be successively decomposed into its cliques.

Let $C_1, C_2, \ldots, C_K$ be a sequence of the cliques in an undirected graph G and define the history and separators, respectively, as

$$H_j = C_1 \cup C_2 \cup \ldots \cup C_j \quad \text{and} \quad S_j = H_{j-1} \cap C_j$$

for $j = 2, \ldots, K$ with $H_1 = C_1$. The sequence is said to obey the running intersection property (RIP) if $S_i \subseteq C_j$ for some $j < i$ for $i = 2, 3, \ldots, K$. The cliques of a decomposable graph can be numbered to have RIP ordering. The cliques C_1, C_2 and C_3 obtained from the graph above,

in that order, are a RIP ordering with separators S_2 and S_3.

Finally, a probability measure can be associated with an undirected graph with each vertex being a random variable. For decomposable graphs, the probability density function can be written in terms of the cliques and separators.

2.2 Contingency tables: notation and model

The outlier detection model described in the article can only be used on the data for which all variables can only take on a finite set of values. Such variables are also called *categorical variables*.

Discrete data set can be appropriately showed by a *contingency table* which summarizes the counts of all combinations of the data set variables. Such way of showing the data is very informative and is a convenient way of describing categorical data sets.

The *dimension* of the table is the number of variables.

Let Δ denote a finite set of discrete variables, in which each variable, $\delta \in \Delta$, takes a value in the level set I_δ. An outcome, $i = (i_\delta)_{\delta \in \Delta}$ is a cell from the set of all cells $I = \times_{\delta \in \Delta} I_\delta$.

The entire contingency table of counts is the set $n = \{n(i)\}_{i \in I}$, where $n(i)$ is the number of observations that falls in cell i and $|n| = \sum_{i \in I} n(i)$ is the total number of observations. The probability that an observation belongs to cell i is denoted as $p(i)$.

For the DGM, probabilities can then be written as

$$p(i) = p_{C_1}(i_{C_1}) \prod_{k=2}^{K} \frac{p_{C_k}(i_{C_k})}{p_{S_k}(i_{S_k})} \tag{1}$$

where $C_1, C_2, \ldots, C_K$ and $S_2, S_3, \ldots, S_K$ are the RIP ordered cliques and separators in a decomposable graph.

2.3 Statistical test for outlyingness

In order to test if a new observation is an outlier, it is assumed that this observation sampled from a distribution different from the distributions of other observations.

A universal definition of an outlier is given by Hawkins (1980): "an observation which deviates so much from the other observations in the data-set as to arouse suspicions that it was generated by a different mechanism." The outlier detection method of [1] directly adapts the definition given by Hawkins by specifying a hypothesis of an outlier being distributed differently than all other observations in a given database.

The null hypothesis is

$$H_0 : q = p$$

where p and q are specified through DGM formula; p is a vector of probabilities for the TRUE distribution, and q is the probability vector of distribution of a new observation. If H_0 is false, the observation is considered an outlier in the table.

Assume that the *likelihood* L_0 expresses how likely it is that z_{new} belongs to the database D. An alternative likelihood L_1 can also be specified, indicating how likely it is that z_{new} does not belong to D. Then the *likelihood ratio* is defined as

$$LR = \frac{L_0}{L_1},$$

which can be shown to be completely specified through the counts of observations in cliques and separators for the given interaction graph. It can be therefore tested if z_{new} is an outlier in D by calculating LR and determining if the value of LR is "too large" in which case it would be rejected that z_{new} comes from D.

For more details, see [1].

3 Description of the simulation plan

The choice of a modelling grid is a difficult one because it is necessary to keep in mind the limited computational resources. Below we describe our simulation plan in terms of various parameters. Its adoption was motivated by the goals of the study and, first of all, by a wish to investigate method's performance on sparse tables.

- *Graphs.* The graphs considered are depicted in figures 2-5. As mentioned previously, by choosing different structures we aimed to investigate whether the table structure defined on the level of DGM affects method's performance.

Figure 2: The simple graph of 2 vertexes and 1 clique

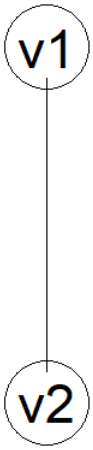


Figure 3: The graph of 9 vertexes and 5 cliques

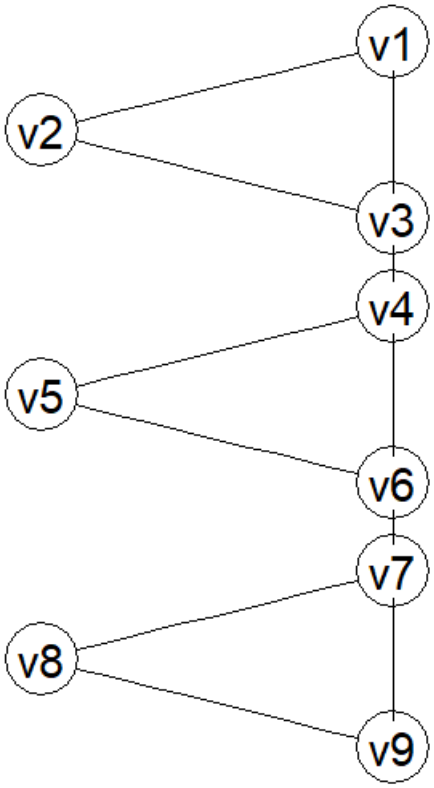
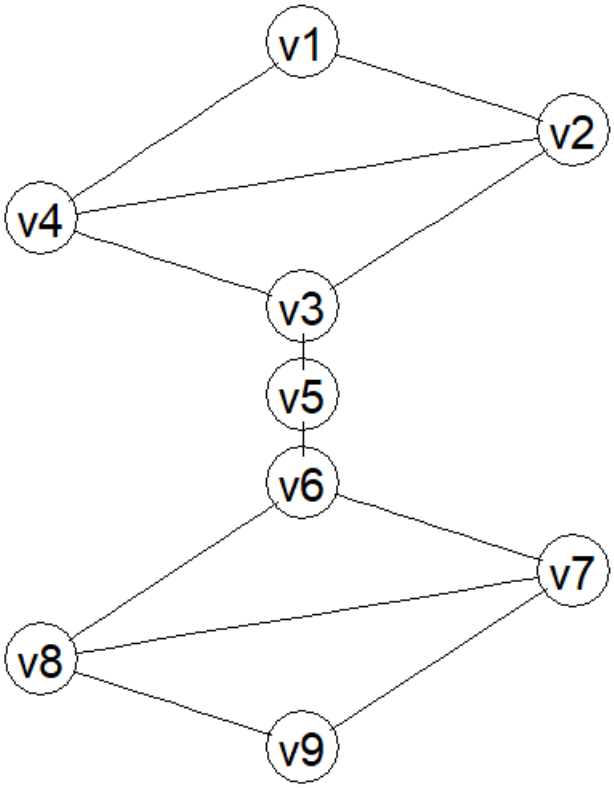


Figure 4: Linear graph of 5 vertexes and 4 cliques

Figure 5: The graph of 9 vertexes and 6 cliques



- The grids.
  - As was described previously, each vertex of a particular graph is a variable of a contingency table which gains certain values. For clearer understanding, for more complex graphs we will describe the size of a contingency table by a number of cells in this table, which is a product of values of all variables in table.
    * For the simplest graph, having only 2 vertexes, we considered 2x2, 5x6 and 9x9 contingency tables.
    * For the second graph of 9 vertexes and 5 cliques contingency tables having 512, 4608, 28800 cells were taken.
    * The linear graph has 5 vertexes and 4 cliques, and for this graph we considered contingency tables of 32, 600 and 5040 cells.
    * The last graph, which is the most complex graph in this work, has 9 vertexes and 6 cliques. We took 512, 6912 and 17280 cells for different models of this graph.
  - *Distributions.* To produce sparse tables, for each DGM and vertex value combination, we considered the following conditional distributions on the cliques (the notions are the same as in Section 2).
    * One parameter $s \in (0,1)$ model used to generate data from tables having TRUE distribution:

$$p(i_{C_1}) = \begin{cases} 1-s, \text{ if } i_{C_1} \text{ is the first value of } C_1; \\ \frac{s}{|C_1|-1} \text{ for other } i_{C_1} \in C_1. \end{cases} \tag{2}$$

$$p(i_{C_k \setminus S_k} \mid i_{S_k}) \overset{k \geq 1}{=} \begin{cases} 1 - s^{\min(5,\, \max\limits_{j \in i_{S_k}} j)}, \text{ if } i_{C_k \setminus S_k} \text{ is the first value of } C_k; \\ \frac{s^{\min(5,\, \max\limits_{j \in i_{S_k}} j)}}{|C_k|-1}, \text{ for other } i_{C_k \setminus S_k} \in C_k \setminus S_k. \end{cases}$$

    * Distributions used generate outliers:
      · the same as in (2) but with value of $s' = s/2$;
      · $i_{C_1} \sim Bin(|C_1|, 0.5)$ and $i_{C_k \setminus S_k} \mid i_{S_k} \sim Bin\left(|C_k \setminus S_k|, 0.5^{\min(5,\, \max\limits_{j \in i_{S_k}} j)}\right)$;
      · i_{C_1} and $i_{C_k \setminus S_k}$ uniform over their domains and independent of values of i_{S_k}.

Note that, our variables (or vertexes in the DGM) attained only positive integer values. Therefore, $\max\limits_{j \in i_{S_k}} j$ was meaningful in our setup. The indices of table values denote the values of its subtables (for example, i_{C_1} denotes the value of the first subtable; $i_{C_2 \setminus S_2}$ denotes the value of the second subtable, from which values of the second separator are excluded). Also, we fixed a lexicographical order on each subtable of the table under consideration. Thus, "the first value" above refers to the table cell having all coordinates equal to 1: $(1, 1, \ldots)$. Taking small s, one ends up with a sparse TRUE table having "large" mass only at some small fraction of the cells. Therefore, in our simulation, TRUE distribution put on the data generating table was always sparse. Outliers' distribution was sparse in the first case. Other choices were for seeing whether higher deviation from the TRUE one improves specificity of

the test. In all cases, in addition to the outliers, TRUE data was also generated in order to investigate method's sensitivity. Values of s ranged over $\{0.3, 0.2, 0.1\}$ for the TRUE distribution (and were taken equal to $s/2$ for the case when outliers were generated by the formula given in (2)). Structural table models (in terms of DGM and vertex specifications) for outliers in all cases were the same.

  - The difference between the TRUE probability distribution and probability distributions of outliers was measured by computing ***Kullback–Leibler divergence*** for each combination of TRUE distribution and outlier distribution.

- *Number of simulations used to estimate type I-II errors* (or sensitivity-specificity) for testing for outliers on one pattern (see Remark 3 below) was taken equal to 100 for the DGM's 2 and 3, and for other two - equal to 500.

- *Number of observations in the TRUE table used to build model for subsequent testing for outliers* on one pattern was tied to the number of cells in the TRUE table and expressed in terms of fractions of this number. By doing so, we aimed to escape the situation when the number of observations is comparable to the number of parameters in the table.
  Let $r = n_{cells}/n_{obs}$, where n_{cells}, n_{obs} stand for the number of cells and number of generated observations respectively. For each pattern, we considered the following fractions expressed as ratios: 0.1, 0.05, 0.01.

In the above, by pattern we mean one model corresponding to the fixed:

1. DGM;
2. numbers of values attained by vertexes;
3. TRUE distribution;
4. outliers' distribution.

□

4 Results

In this section, by making use of operational characteristics of the method, we describe the results obtained, which allow us to draw conclusions about the performance of the method. The investigated operational characteristics of the method are defined through conditional probabilities. We use such definitions of ***sensitivity*** and ***specificity***:

sensitivity $= \mathbb{P}$(method classified observation as non-outlier | observation is really not outlier);

specificity $= \mathbb{P}$(method classified observation as outlier | observation is indeed a real outlier).

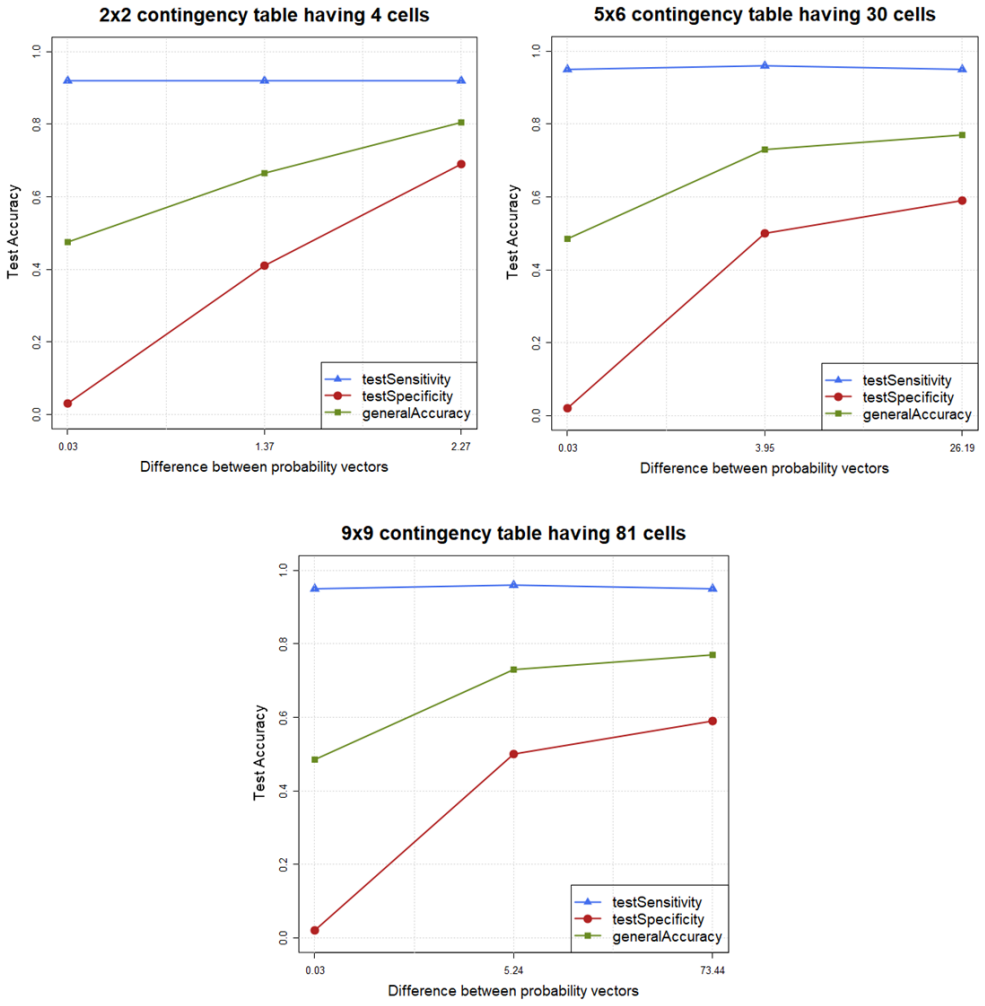
Let's move on to the description of the obtained results.

4.1 When model assumptions are met

4.1.1 How characteristics of the test change when the distribution used to generate outliers become more and more different from the distribution used to generate TRUE observations?

1. For the first 100 simulations we took the grids of a simplest graph (see description of a simulation plan 3 above) and generated outliers and TRUE observations according to the distributions described in plan (3). The ratio to calculate number of observations in the TRUE table was always fixed in the case of analyzing the dependency on different distributions for outliers, and was taken equal to 0.01.

Figure 6: Graphics of test accuracy for the simplest models: dependence on difference between probability distributions

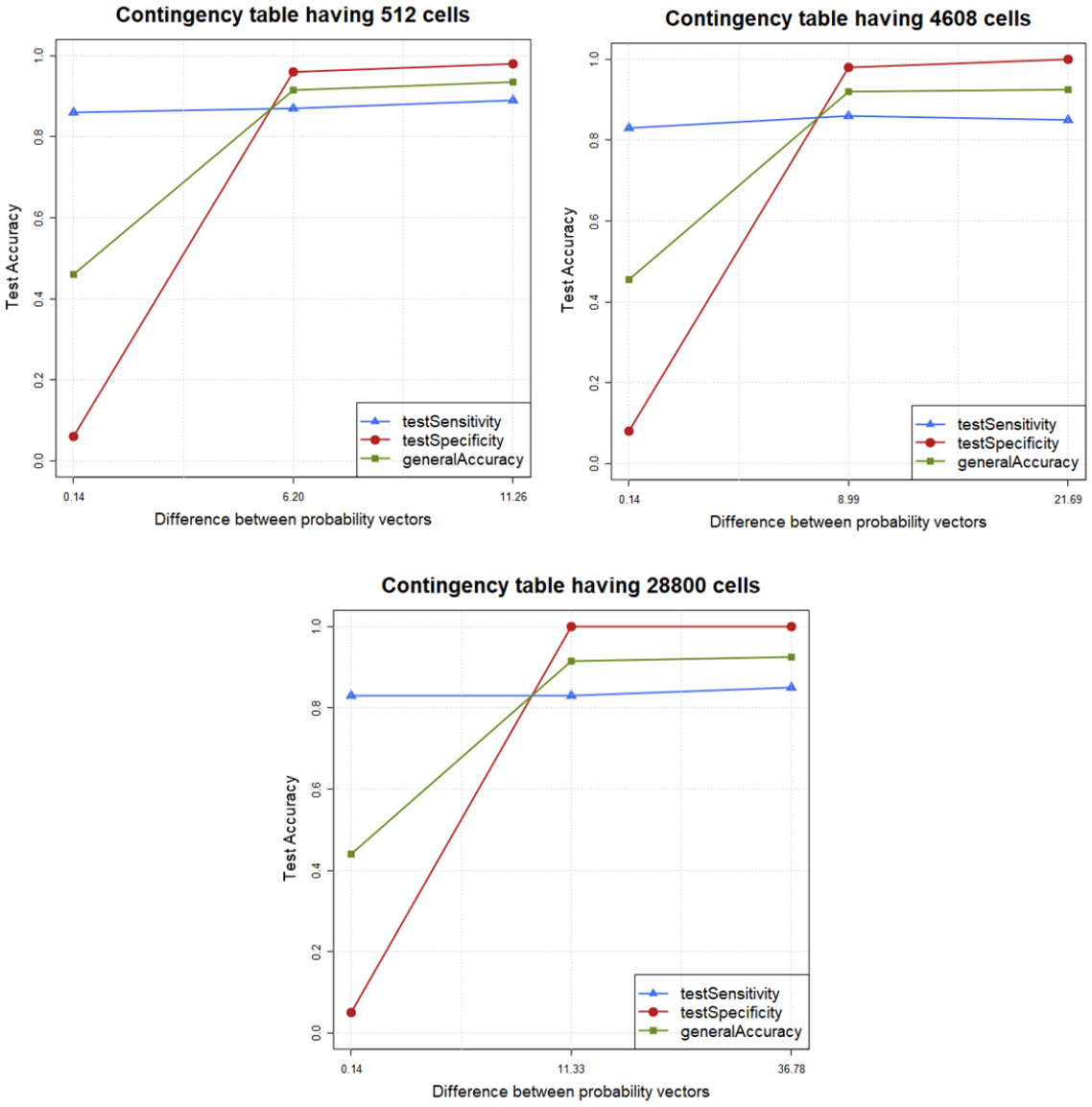


From the graphics above we can see that test sensitivities are high, while test specificity grows as the distance between probability vectors become bigger. In the first case, when

distribution of outlier is very similar to the TRUE distribution, test specificity is very low. Overall accuracy is not so good because of the test specificity, it changes from 0.45 to 0.80.

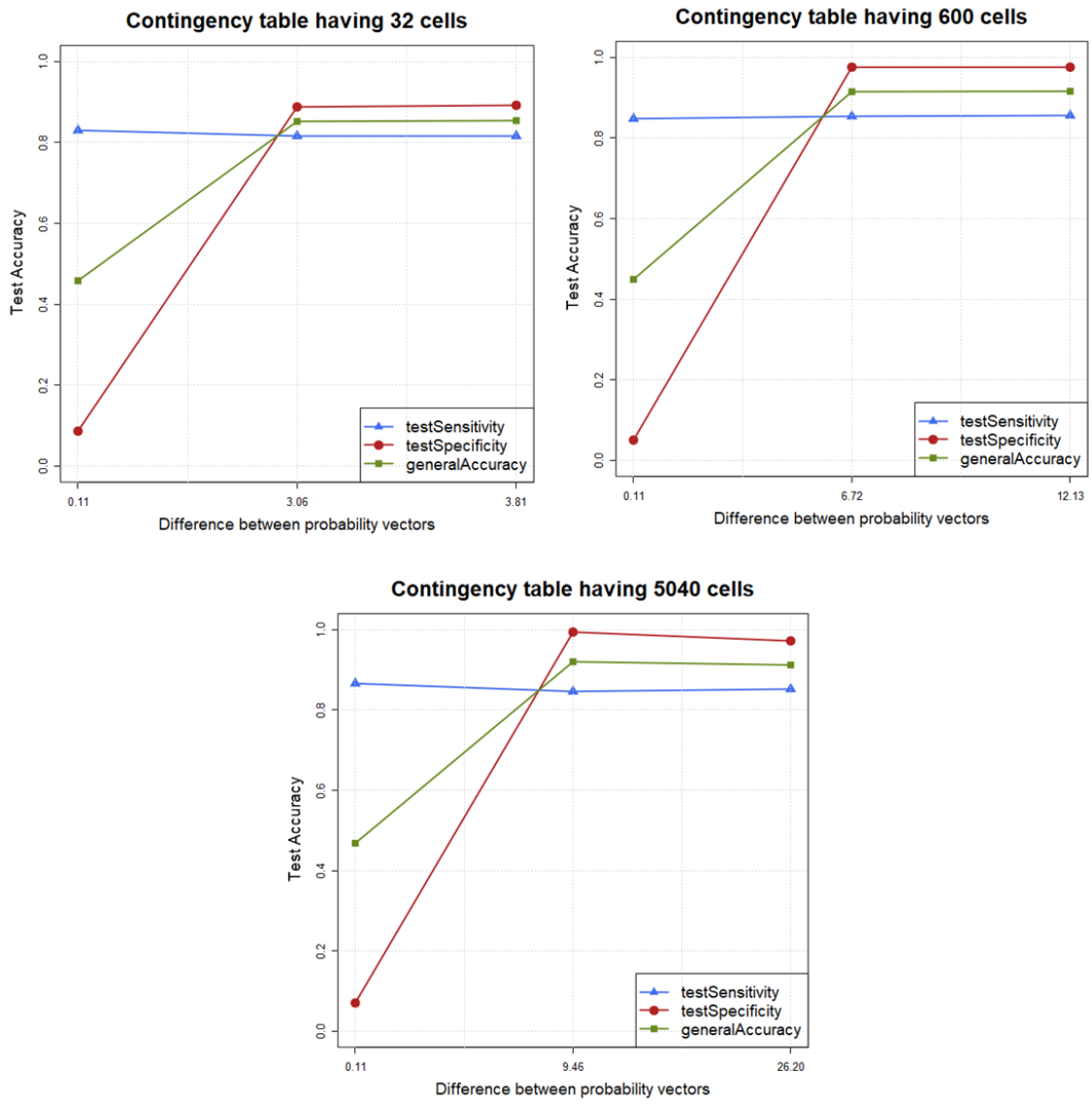
2. Let's take now the grids of a complex graph of 9 vertexes and 5 cliques (see Figure 3). Here test sensitivity also remains good (around 80 and 90 percent), test specificity jumps up when the distribution for generating outliers becomes not the same as for generating TRUE observations. The general accuracy also improves when the distribution of outliers changes. Overall, we can notice that test characteristics are better for the more complex models.

Figure 7: Graphics of test accuracy for the models of a second DGM: dependence on difference between probability distributions



3. For linear graph patterns we used 500 simulations as was defined in previous section (see 3). The ratio to calculate number of observations in TRUE table was always fixed and equal to 0.01.

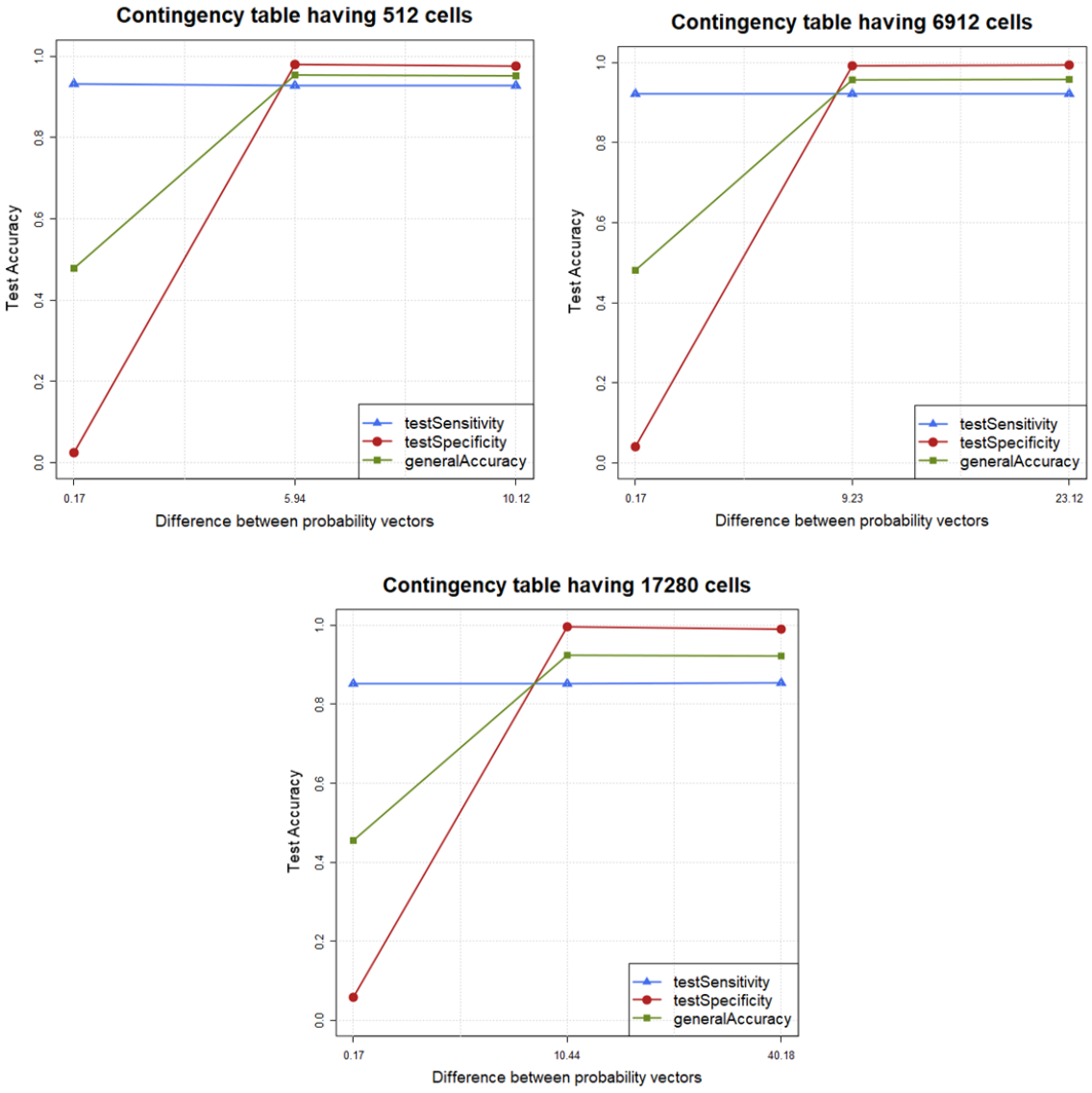
Figure 8: Graphics of test accuracy for patterns of linear graph: dependence on difference between probability distributions



In graphics above we see that sensitivity of the method is always quite high (80-85%), test specificity is also high when distributions of outliers are more different from the TRUE distribution. What is more, it is noticeable that test sensitivity slightly improves when number of cells in contingency table becomes bigger, and specificity of the method rises by 9-10% as well.

4. For the last and the most complex graph patterns we used 500 simulations as well. The ratio to calculate the number of observations in the TRUE table was fixed and equal to 0.1.

Figure 9: Graphics of test accuracy for patterns of the most complex graph: dependence on difference between probability distributions

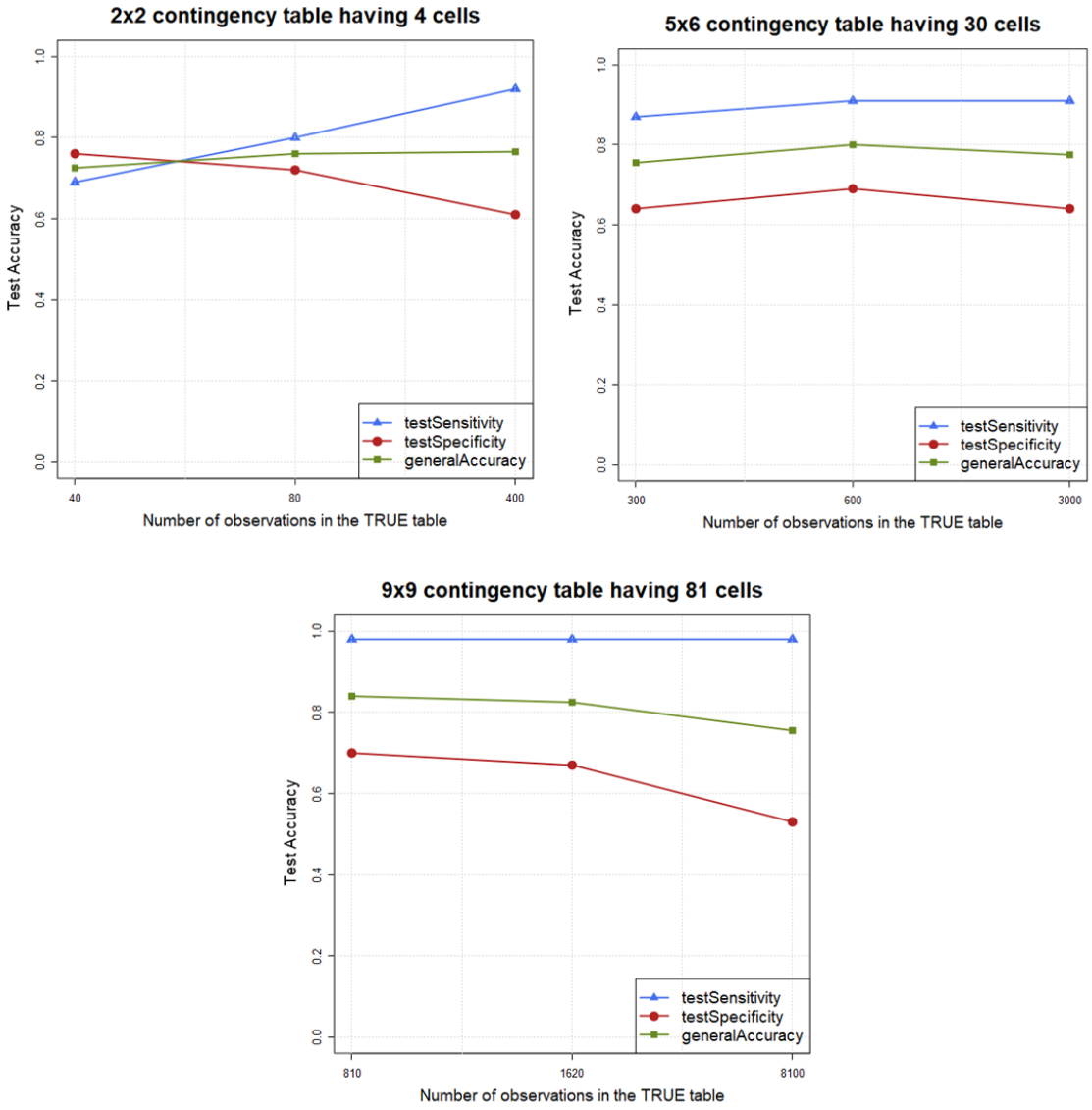


It is visible that sensitivity of the method becomes a little bit worse with the growth of number of cells in table, but still remains higher than 80%. Test specificity is always higher than 90% for the distributions of outliers that sharply differ from the TRUE distribution.

4.1.2 How characteristics of the test change when number of observations increases in TRUE table used to build the model for testing?

1. Here we start from the simplest graph again and take its grids (2). We fix only one distribution used to generate outliers in this case - the most different from TRUE distribution. The ratios to calculate number of observations in TRUE table were taken equal to 0.1, 0.05, 0.01 in this case of testing the dependency on number of observations in table. For each pattern (see Remark 3) we take number of simulations equal to 100.

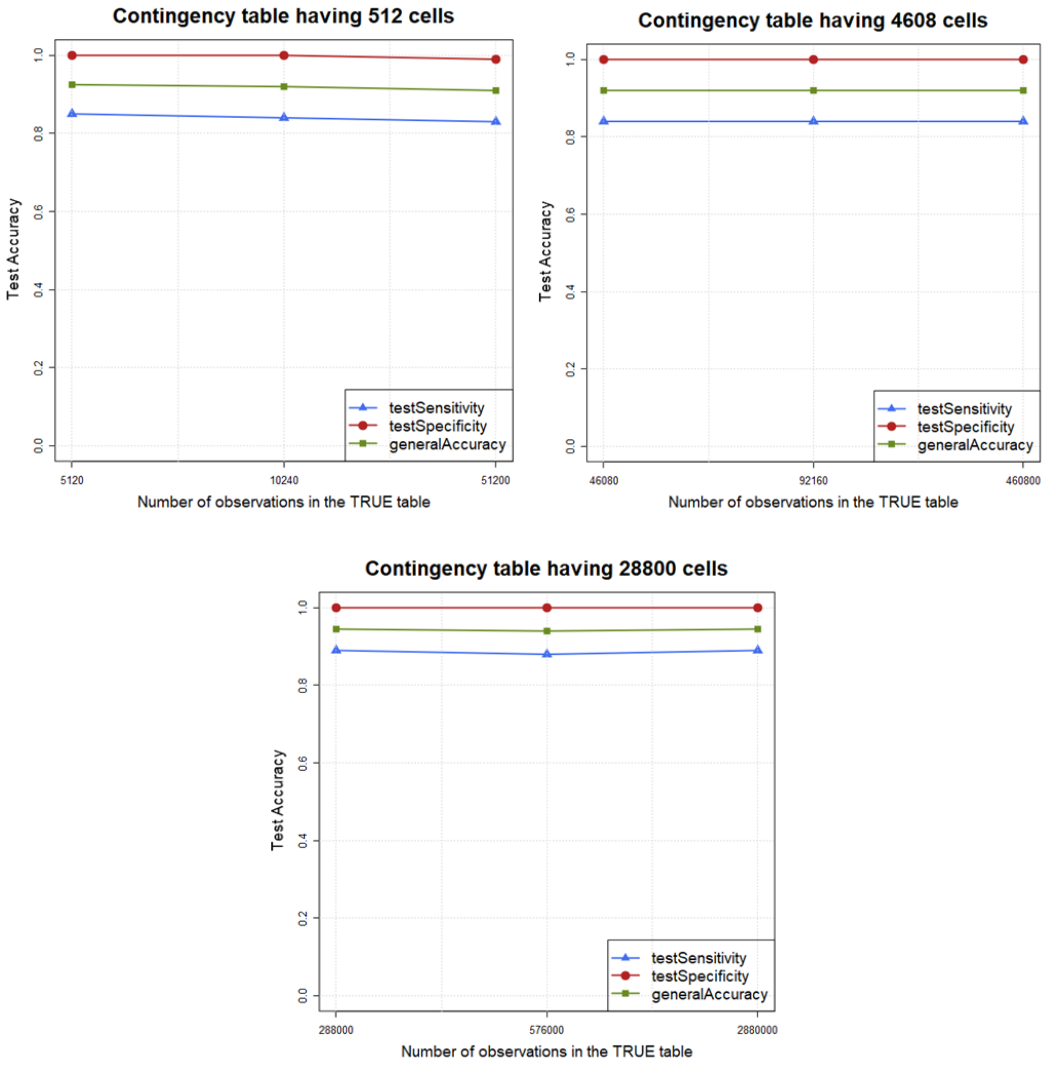
Figure 10: Graphics of test accuracy for the simplest models: dependence on number of observations in TRUE table



From the graphics it can be seen that test sensitivity grows with the number of observations or stay always high as in third graphic, but test specificity decreases when the number of observations grows, which was not expected. The general accuracy varies around 80%.

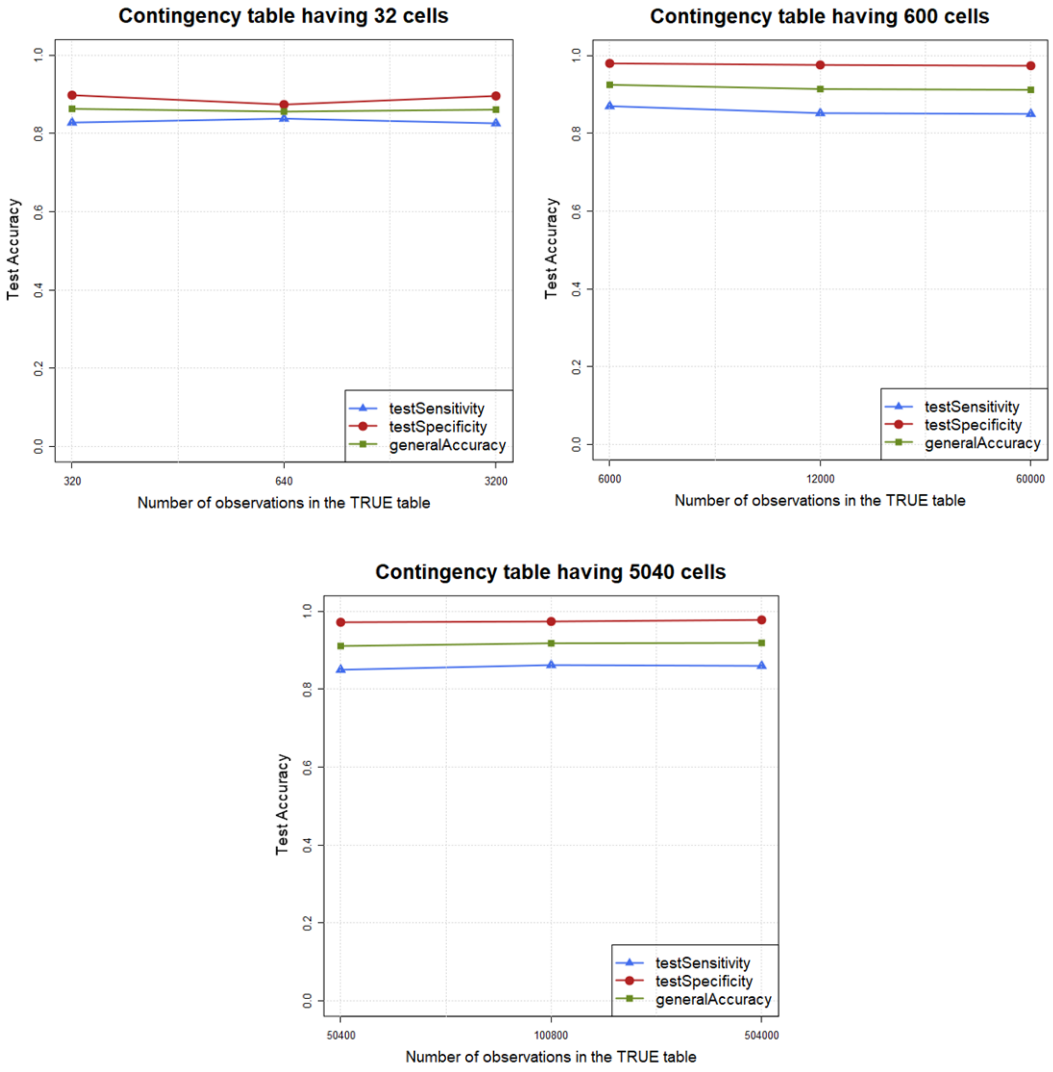
2. Let's take now the grids of a complex graph of 9 vertexes and 5 cliques (3). Here we see that test specificity is perfect and almost always one hundred percent, test sensitivity varies around 80-90%. General accuracy is also very high.

Figure 11: Graphics of test accuracy for the models of a second DGM: dependence on number of observations in TRUE table



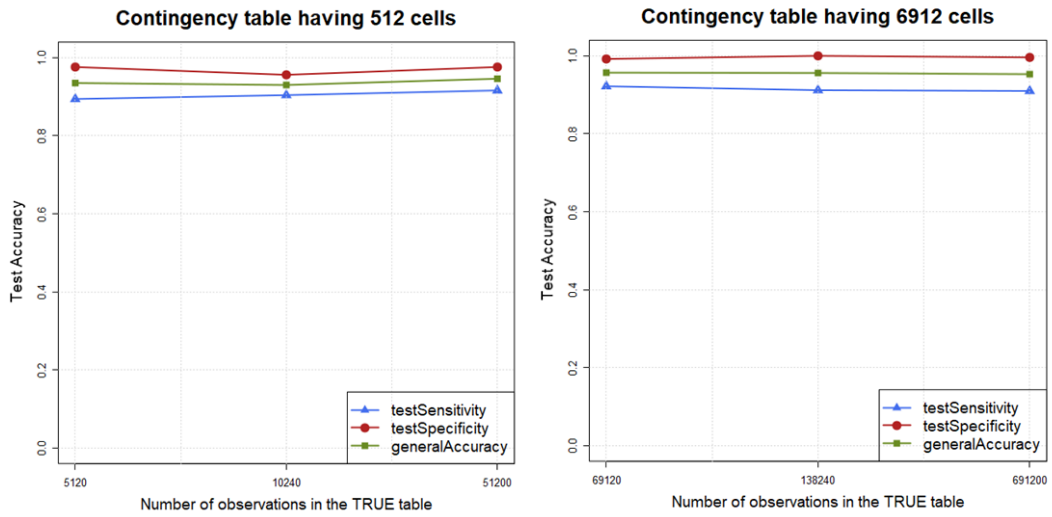
3. For linear graph models (5 vertexes and 4 cliques) test specificity is much higher for the tables with bigger number of cells, test sensitivity looks very similar in all cases. There is not seen any improvement in method's performance with the growth of number of observations in TRUE table.

Figure 12: Graphics of test accuracy for the models of a linear graph: dependence on number of observations in TRUE table



4. Unfortunately, as the last graph was the most complex graph with 9 vertexes and 6 cliques, and the number of simulations was quite large - equal to 500, we ran out of computational resources and therefore we were unable to test a table with a very large number of cells. However, we still can draw conclusions from two graphics below:

Figure 13: Graphics of test accuracy for the models of the most complex graph: dependence on number of observations in TRUE table



We see that test specificity is very high; test sensitivity in the first case increases with the larger number of observations in TRUE table, but in second case it decreases a little bit. General accuracy holds higher than 90% all the time.

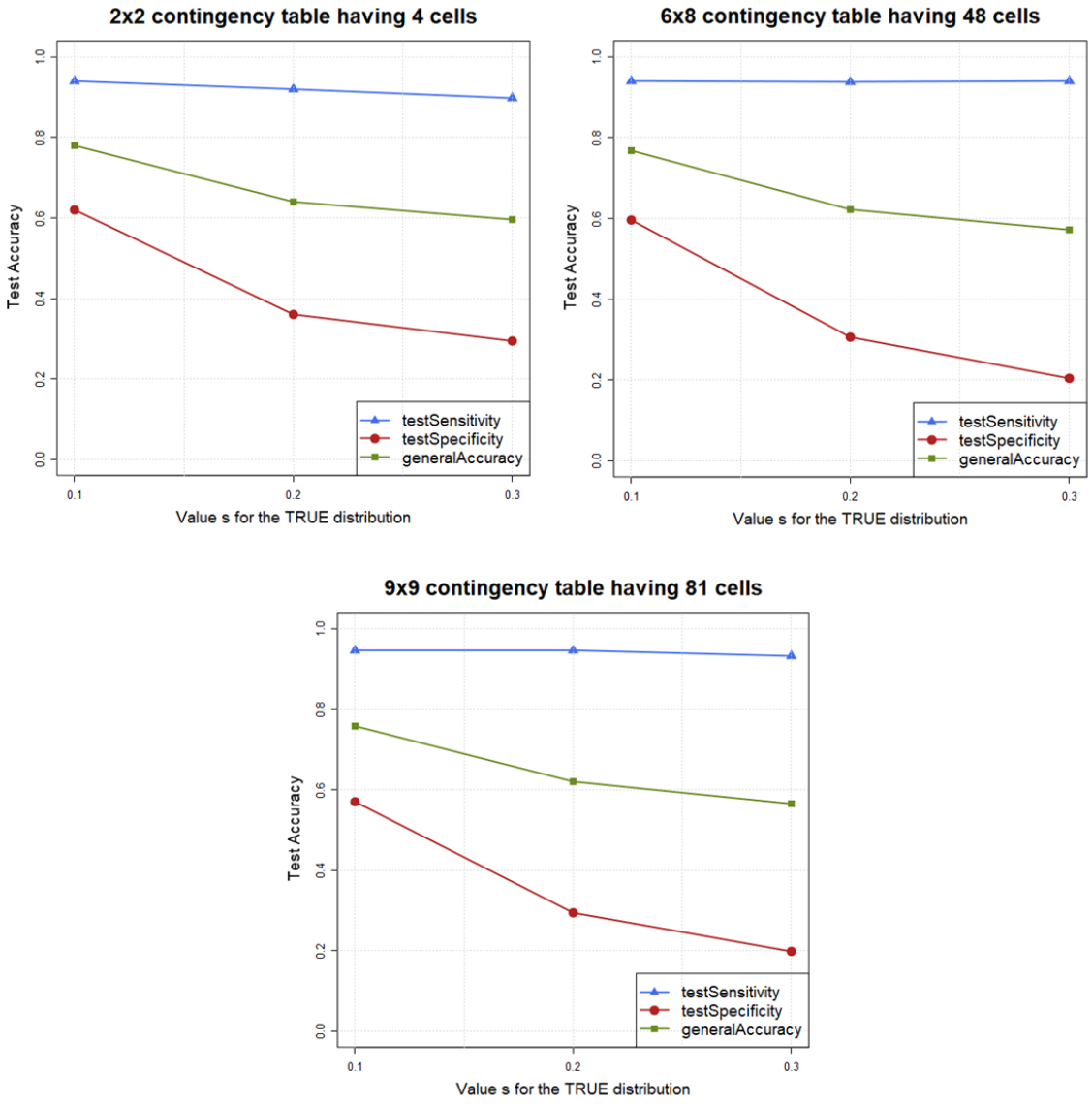
Overall, there is not seen a tendency of method's performance improving with the higher number of observations in table. However, test characteristics for the most cases, particularly for the cases of more complex graphs, are always good.

4.1.3 How characteristics of the test change when values of s range over {0.3, 0.2, 0.1} for the TRUE distribution?

Remark: parameter ***s*** *is responsible for the sparsity of the table generated.*

1. To begin with, let's take models of the simplest graph of two vertexes and one clique. We fix the binomial distribution for outliers (see description in section 3), as it is the most different from the sparse TRUE distribution and gives the highest specificity. The ratio for calculating number of observations in TRUE table is also fixed and equal to 0.01. Number of simulations here is taken equal to 500.

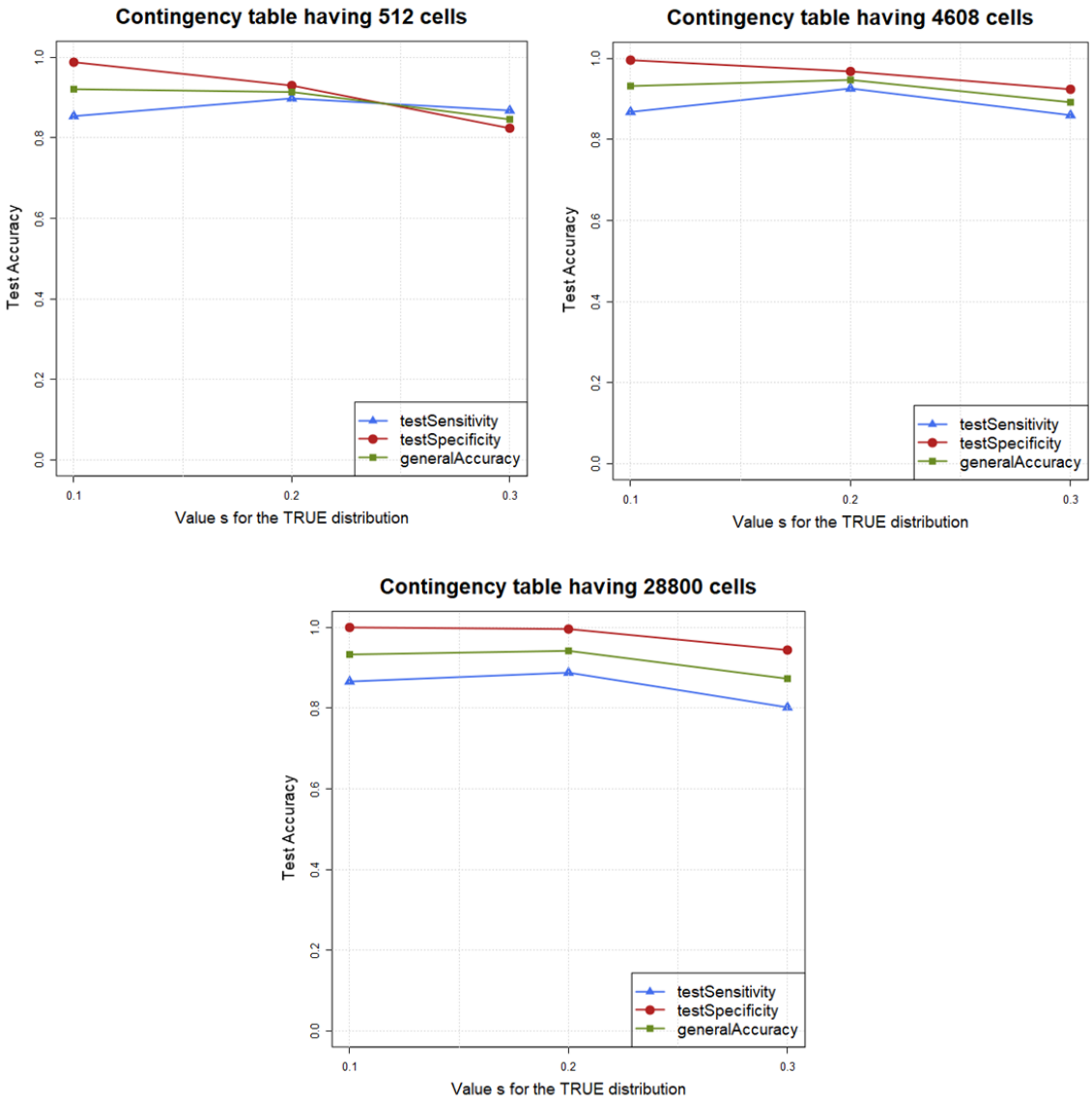
Figure 14: Graphics of test accuracy for models of the simplest graph: dependence on sparsity of TRUE table



From figures above we can clearly see that with the higher value of s of TRUE distribution performance becomes worse. However, despite the fact that test specificity drops dramatically, test sensitivity stays quite high, which means that method recognizes TRUE observations quite well.

2. Taking models of second DGM we also fix binomial distribution for outliers, but ratio for calculating number of observations in TRUE table is equal to 0.1 in this case. Number of simulations is taken equal to 500.

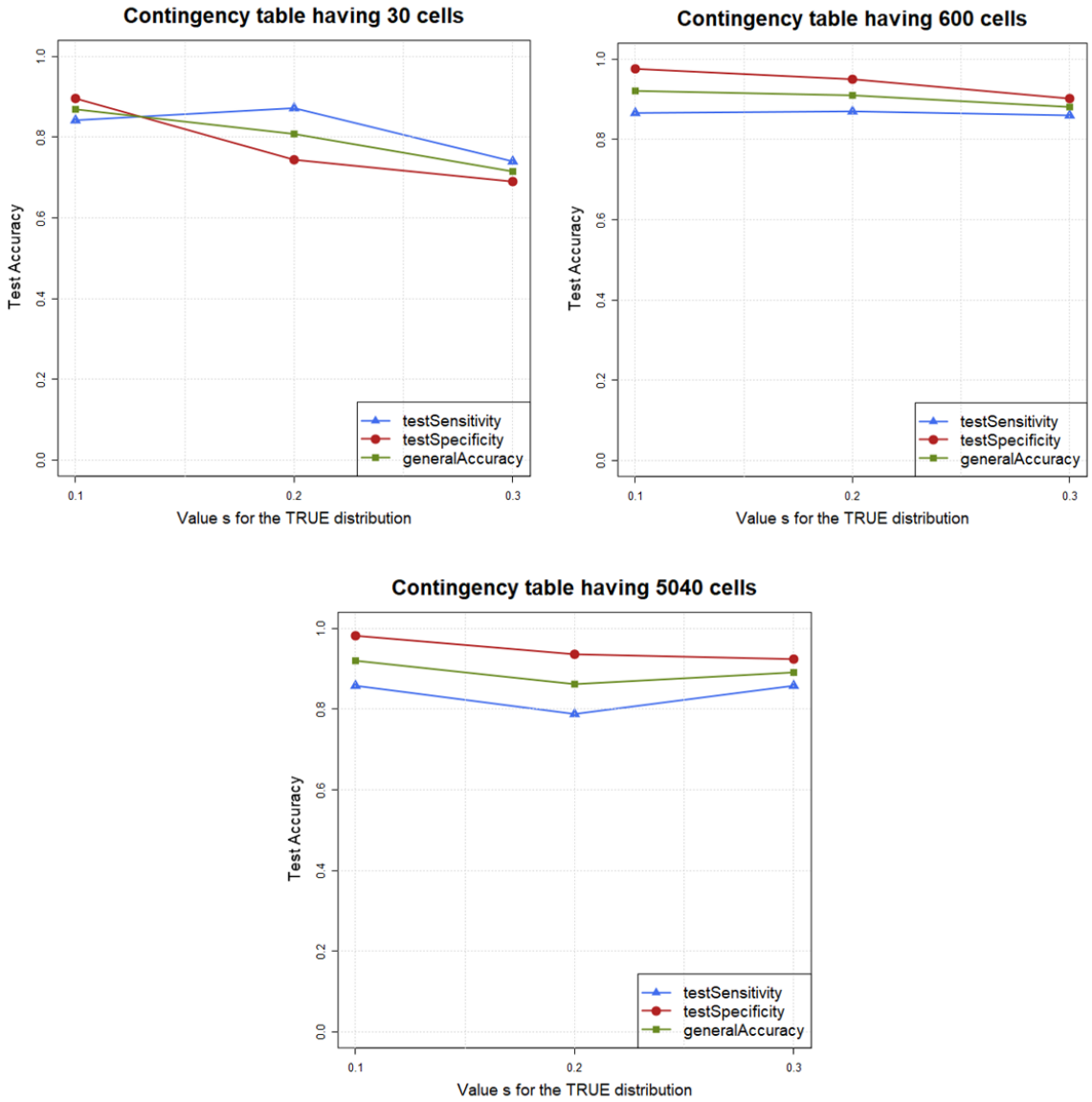
Figure 15: Graphics of test accuracy for models of the second graph: dependence on sparsity of TRUE table



The general performance is much better in the case of more complex graph comparing to the models of the simplest graph. Test specificity is highest when distribution of TRUE tables is the most sparse; then it starts to worsen. Test sensitivity varies around 80-90%. Overall, we see that the highest results of test specificity are obtained when we have the largest number of cells in contingency table.

3. For linear graph models we take ratio for calculation of number of observations in TRUE table equal to 0.01 again, as linear graph models are simpler than models of the second graph and our computational resources allow us to take larger number of observations for models of this graph. Number of simulations is taken equal to 500.

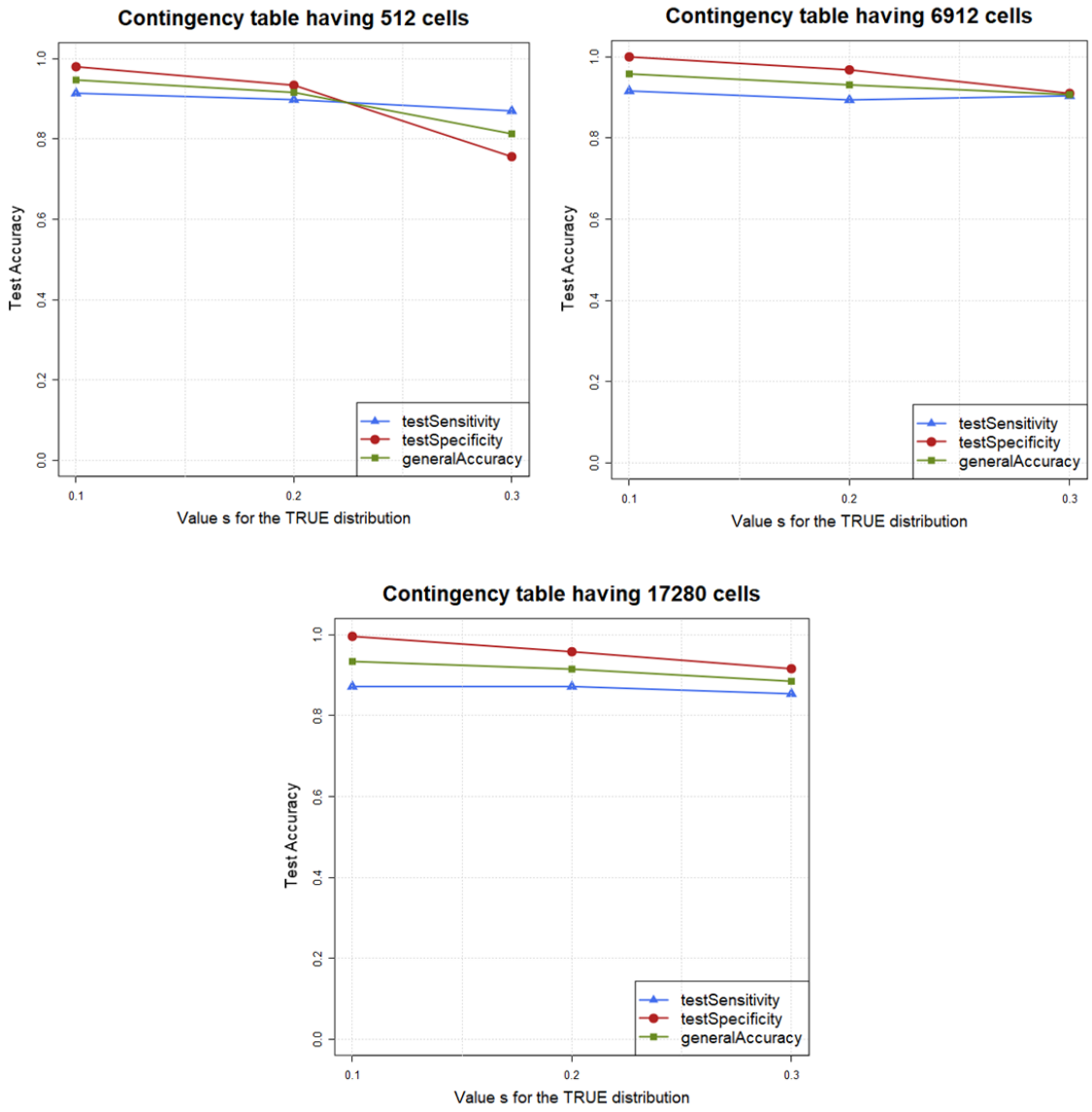
Figure 16: Graphics of test accuracy for models of the linear graph: dependence on sparsity of TRUE table



Test specificity in the first case is not as high as it is in next two cases - probably because of the small number of cells in contingency table; when value of parameter s increases, test specificity becomes even worse. However, it stays quite high in next two cases. Test sensitivity does not reach 90% in any case, but we see that in some cases it improves with the larger value of s, whereas in other - worsens or stays the same.

4. For the last and the most complex graph models we take the ratio for calculating the number of observations in TRUE table equal to 0.1 and number of simulations equal to 500.

Figure 17: Graphics of test accuracy for models of the most complex graph: dependence on sparsity of TRUE table



Specificity in general is much higher comparing to linear graph models, but it also worsens when contingency table becomes less sparse; test sensitivity is higher as well. General accuracy is the best for the most complex graph models.

4.2 When model assumptions are not met

In this subsection we aimed to look what happens to method's performance when some of the model assumptions under which it was developed are violated. The model described in [1] needs to be *saturated*. Cell probabilities of a *saturated* model are not restricted in any way, except for the constraints of being ***positive*** and summing to one. We met these assumptions earlier in our work, as well as other assumptions of the model described in [1], but in this subsection we will violate the condition of *saturated* model. For TRUE table and outliers we created sparse distributions, cell probabilities of which include zeros. Let's take a look at the performance of the method, when one condition of a saturated model is violated.

- – Consider the following conditional distributions on cliques for the sparse TRUE table ($s = 0.05$):

$$p(i_{C_1}) = \begin{cases} 1-s, \text{ if } i_{C_1} \text{ is the first value of } C_1; \\ s, \text{ if } i_{C_1} \text{ is the second value of } C_1; \\ 0, \text{ for other } i_{C_1} \in C_1. \end{cases} \tag{3}$$

$$p(i_{C_k \setminus S_k} \mid i_{S_k}) \overset{k>1}{=} \begin{cases} 1 - s^{\min(5, \max\limits_{j \in i_{S_k}} j)}, \text{ if } i_{C_k \setminus S_k} \text{ is the first value of } C_k; \\ s^{\min(5, \max\limits_{j \in i_{S_k}} j)}, \text{ if } i_{C_k \setminus S_k} \text{ is the second value of } C_k; \\ 0, \text{ for other } i_{C_k \setminus S_k} \in C_k \setminus S_k. \end{cases}$$

  – Distribution used to generate outliers:

$$p(i_{C_1}) = \begin{cases} s, \text{ if } i_{C_1} \text{ is the first value of } C_1; \\ 0, \text{ for other values } i_{C_1} \in C_1 \text{ except the last value of } C_1; \\ 1-s, \text{ if } i_{C_1} \text{ is the last value of } C_1. \end{cases} \tag{4}$$

$$p(i_{C_k \setminus S_k} \mid i_{S_k}) \overset{k>1}{=} \begin{cases} s^{\min(5, \max\limits_{j \in i_{S_k}} j)}, \text{ if } i_{C_k \setminus S_k} \text{ is the first value of } C_k; \\ 0, \text{ for other values } i_{C_k \setminus S_k} \in C_k \setminus S_k \text{ except the last value of } C_k; \\ 1 - s^{\min(5, \max\limits_{j \in i_{S_k}} j)}, \text{ if } i_{C_k \setminus S_k} \text{ is the last value of } C_k. \end{cases}$$

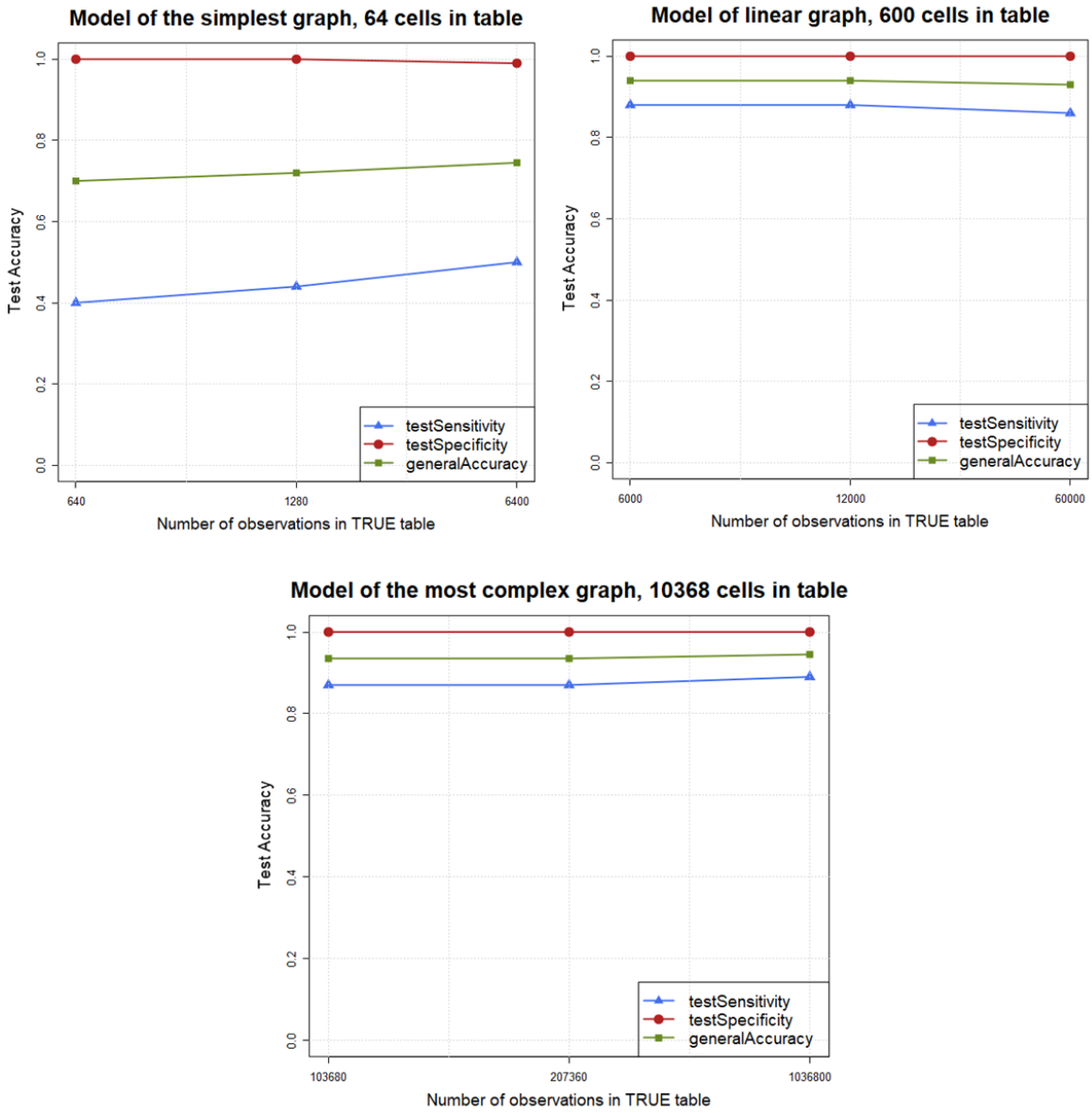
Note that here our variables (or vertexes in the DGM) attained only positive integer values. Therefore, $\max\limits_{j \in i_{S_k}} j$ is meaningful in our setup. Taking small s ($s = 0.05$), we have a sparse TRUE table having "large" mass only at some small fraction of the cells. Outliers' distribution is also sparse, but it differs significantly from the distribution of TRUE table (the distance between probability vectors was measured by computing ***Kullback–Leibler divergence***). Structural table models in terms of DGM and vertex specifications are the same as described in simulation plan (3).

Three graph structures for this analysis were taken from section (3): the simplest graph, linear graph and the most complex graph (figures 2, 4 and 5). For each DGM we considered only one vertex value combination, so we have 3 models here, each of which has

different graph structure. Number of simulations used to estimate sensitivity and specificity for testing on one pattern was taken equal to 100. The ratios to calculate number of observations in TRUE table were taken equal to 0.1, 0.05, 0.01, as we used previously in our work.

Thus we obtain the following results:

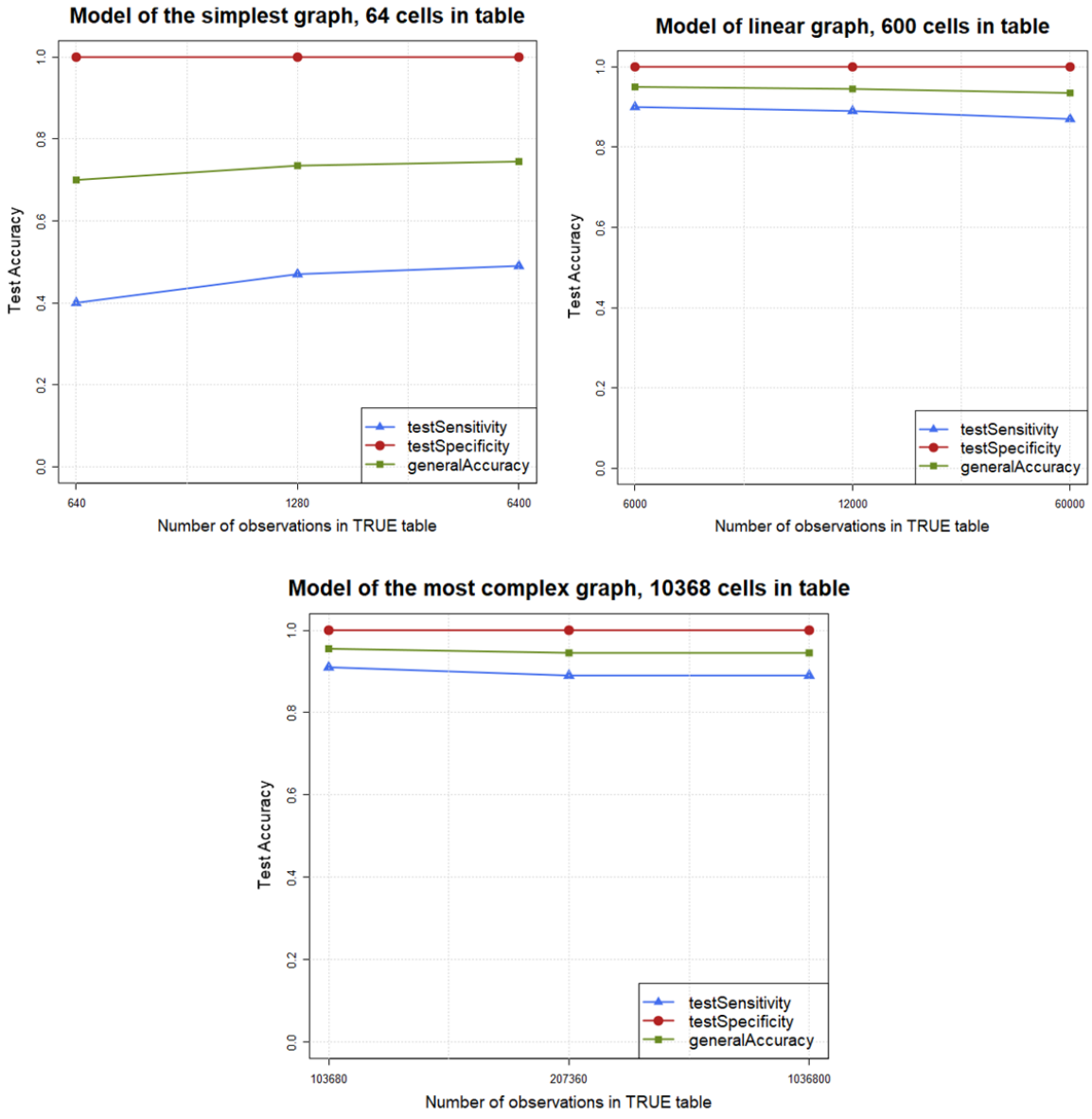
Figure 18: When probability distributions of TRUE table and outliers include zeros



From the figures above we can notice, that test specificity is always very high for all models of all graph structures. However, for the model of the simplest graph test sensitivity is quite poor. Test sensitivities for more complex models (of more vertexes and cliques) are high enough, as it was in cases described previously in our work (when all assumptions of the model were met).

- Let's try to look at model's performance, when distribution of outliers is binomial (see plan description 3). The distribution of TRUE table will be as described in formula (3) of this subsection. Number of simulations and ratios to calculate the number of observations in TRUE table are the same.

Figure 19: When only the probability distribution of TRUE table include zeros

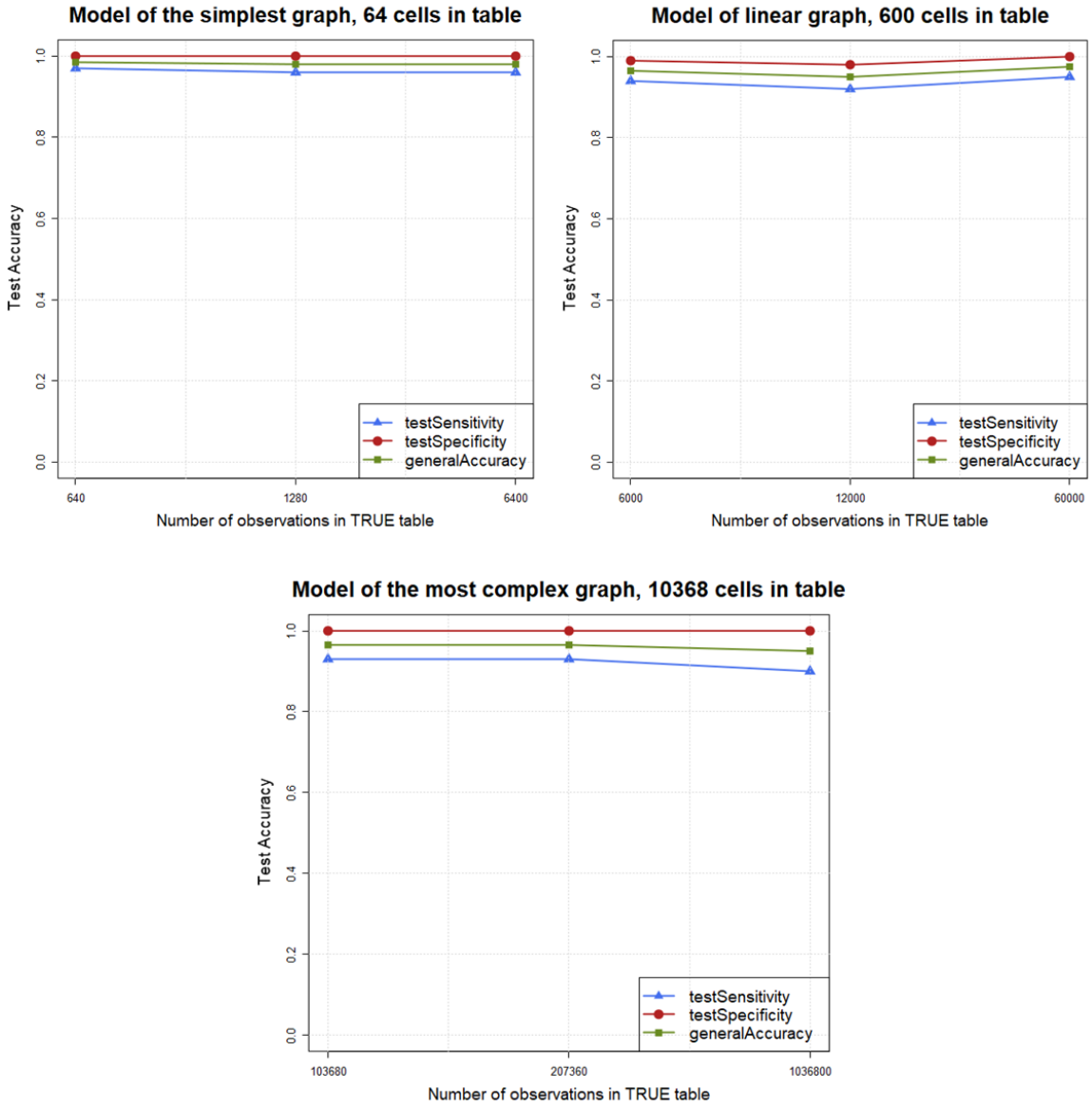


Results are very similar to previously obtained results. All characteristics for models of more complex graphs are good, except sensitivity of the pattern of a simple graph, which varies around 40-50%. We see that zero probabilities in TRUE table affect only sensitivity of the test for pattern of simple graph (recall that test sensitivities where always 80-90% for all patterns of all graph structures, when model assumptions were fully met). The difference between distribution of TRUE table and both sparse and binomial outlier distributions is quite large, so this may be the reason for specificity staying always very

high.

- In the last example we decided to assign binomial distribution to the TRUE table, and sparse distribution with zero probabilities (formula (4)) - to distribution of outliers. We were curious to look if the right distribution of TRUE table improves sensitivity of the test of a simple graph model in spite of the fact that distribution of outliers includes zero cell probabilities.

Figure 20: When only outliers' probability distribution include zeros



Our expectations were met - sensitivity for the model of a simple graph improved dramatically. All characteristics now are very good for all models.

Overall, the conclusion for this subsection would be the following: violation of one condition of saturated model affects only patterns with the simplest graph structure, where we have only one clique of the graph and two vertexes. It mainly affects sensitivity of the test, because specificity mostly depends on difference between probability distributions of TRUE observations and outliers, which is quite large in our examples.

5 Discussion

- Our main goal of the thesis was to estimate type I-II errors of the new method of [1] by making a simulational study and evaluate how performance of the method depends on configuration of different tables being tested. The data in the form of contingency tables was generated according to the described model of the article. All assumptions of the model under which it was developed were taken into account, as well as other conditions and directions of the used method were met. The whole work done and results obtained allow to make some insights:
  - distribution of outlier observations is very important: seeking for better performance of the method, outlier distribution needs to be as much different from the TRUE distribution as possible;
  - number of observations in TRUE table is not as significant as it was expected: sometimes, the smallest number of observations of three taken showed the best performance of sensitivity and specificity. However, it is not advisable to take smaller number of observations than ratio of number of cells and number of observations equal to 0.1;
  - The sparsity of TRUE table used to build model for subsequent testing for outliers quite strongly influences the performance of the method: the bigger the value of parameter s, the more inferior the performance of the method, as the larger value of s determines less sparse distribution in TRUE table, which makes TRUE distribution become more similar to other distributions and worsens the goodness of the test.

  Overall, as we analysed the performance dependency on different graph structures and combinations of vertex values for different graphs, our investigation lets us state that taking complex graph structures appear in better results of test characteristics. We were analysing 4 graph structures, 3 of which were complex; however, we can not affirm that models of a complex graph with a little bigger number of cliques and vertexes show appreciably better performance than any one model of a complex graph with fewer cliques and vertexes. The largest number of cells in contingency table did not improve significantly performance of any pattern of particular graph structure as well; we saw improvements only in some cases, mostly when analysing dependency on sparsity of a TRUE table. However, there is a meaningful difference between the performance of models of simple graph structure and models of complex graph structures. The reason for such distinction could be the fact that the method described and the model proposed in article were mainly adapted to contingency tables of higher dimensions, which can be created only through complex graph structures.

- Unfortunately, our study was limited by the lack of computational resources. All simulations were done using programming language **R**, which has limited memory. The lack of

memory hindered usage of larger number of observations in simulations, which could have further improved performance of the method. What is more, computational resources did not allow to choose larger number of configurations, as well as limited number of cores in computer's processor, which was followed by limited computational speed, prevented from taking more than 500 simulations for one pattern.

Nevertheless, we are satisfied with our investigation and obtained results, as they provided with more experience and knowledge, which will be beneficial in future researches.

References

[1] Mads Lindskou, Poul Svante Eriksen, and Torben Tvedebrink. Outlier detection in contingency tables using decomposable graphical models. *Scandinavian Journal of Statistics*, 47(2):347-360, 2020.

[2] Kuhnt, S., Rapallo, F., Rehage, A. (2014). Outlier detection in contingency tables based on minimal patterns. *Statistics and Computing, 24*, 481–491.

[3] Rapallo, Fabio. Outliers and Patterns of Outliers in Contingency Tables with Algebraic Statistics. *Scandinavian Journal of Statistics 39*, no. 4 (2012): 784-97.

[4] Yick, John S. and Lee, Andy H., (1998). Unmasking outliers in two-way contingency tables. Computational Statistics and Data Analysis, 29, issue 1, p. 69-79.